**Aalborg Universitet**

AALBORG UNIVERSITY
DENMARK

# Condenser

*a platform for handling quantitative proteomic data from matrix science's mascot distiller*

Stensballe, Allan; Kjeldal, Henrik; Knudsen, Anders Dahl

*Publication date:*
2011

*Document Version*
Early version, also known as pre-print

Link to publication from Aalborg University

*Citation for published version (APA):*
Stensballe, A., Kjeldal, H., & Knudsen, A. D. (2011). *Condenser: a platform for handling quantitative proteomic data from matrix science's mascot distiller*. Poster presented at 2nd Nordic Proteomics Symposium 2011, Copenhagen, Denmark.

**General rights**
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

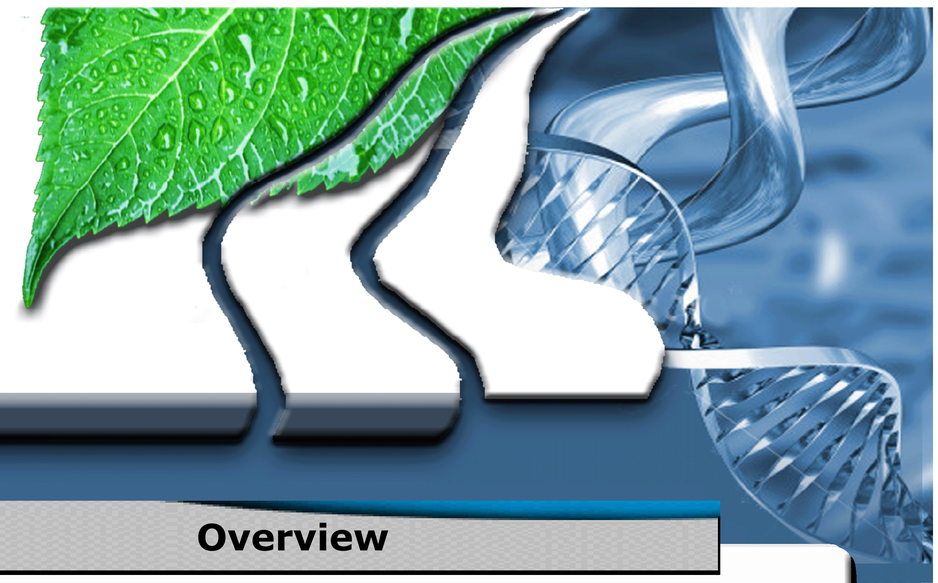
**Take down policy**
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: May 06, 2023

# CONDENSER: A PLATFORM FOR HANDLING QUANTITATIVE PROTEOMIC DATA OUTPUT FROM MATRIX SCIENCE'S MASCOT DISTILLER™

***Anders Dahl Knudsen, H. Kjeldahl, D. Otzen and Allan Stensballe***

*Section of Biotechnology, Department of Biotechnology, Chemistry and Environmental Engineering, Aalborg University; email: as@bio.aau.dk or adk@bio.aau.dk*



## Abstract

- **Condenser is a novel computational tool for aggregating and merging quantitative xml outputs from the Matrixscience Distiller work package into a common format ready for subsequent bioinformatics' investigations.**
- **Condenser allows the quantitative information extracted from the individual sample files using Mascot Distiller into an SQL database. This allows peptides of the same protein accession number in separate samples to be aggregated to a condensed protein list with accurate protein quantifications.**

## Introduction to Condenser

Isotopic mass tag methods have allowed mass spectrometry-based proteomics to overcome problems with variable analyte ionization efficiency and turn quantitative. The nature of quantitative information in mass spectra generally require a detailed evaluation of profile spectra and chromatograms not necessary for database search algorithms that essentially just require peak lists with m/z values and intensities. The development of such tools has been slow and fragmented, with most tools aimed at specific workflows or restricted to certain instrument vendor platforms.

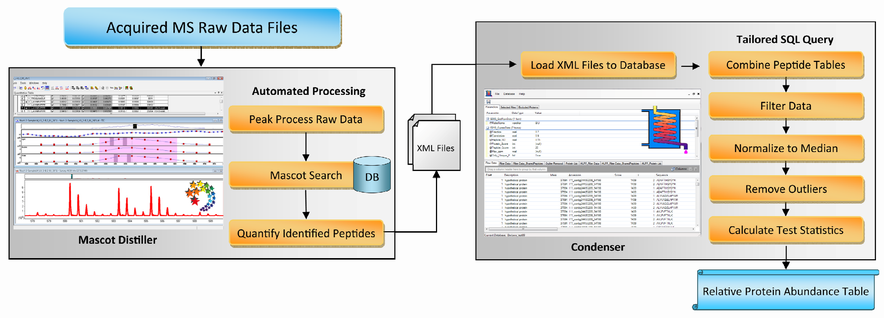
Subsequently, large parts of the community have had limited access to robust computational and statistical tools for quantitative experiments. Matrix Science's Mascot Distiller support multiple labelling approaches and data formats, while being user friendly, however, an unfortunate drawback of Distiller is the lack of sufficient multidimensional experiment support. Condenser appends this functionality to the Distiller workflow, allowing proteins and peptides originating from sequential MS analyses (gel bands, SCX fractions etc.) to be aggregated into combined lists, with statistical evaluations and re-calculation of protein abundance ratios. Loading data into a local database, Condenser processes the data using tailored queries, which allows significant freedom in generating output tables compatible with downstream applications such as repositories, pathway analyses or Gene Ontology annotation tools. However, the solution presented here requires access to a workstation with Mascot Distiller

## Overview

Matrix science's Distiller handles most known MS quantization methods and vendor data formats in terms of relative quantitation of analytes in one sample compared to another.

To adapt Distillers open-ended quantitative workflow to support shotgun proteomic experiments (e.g. GeLC-MS/MS or MudPIT) we have developed a freely available SQL loader application that parses Distiller XML files into a local SQL database. We have established an optimized query for aggregating sequential analyses into a single list of relative or absolute protein abundances complete with integrative statistical testing.

The Distiller XML files are parsed by Condenser into a user-specified SQL database. Individual databases can be created for projects, replicates, etc. or general databases can be created and sub-queried using limiting statements. Condenser uses queries to perform the aggregation, filtration and normalisation of the data and finally generate a list of proteins with appropriate test statistics allowing the user to gauge the significance of the observed changes in protein abundances.



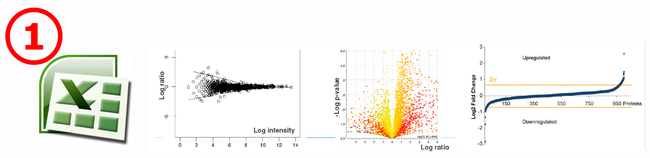

Because of the nature of the resulting ratios all statistical calculations are performed in log -space to ensure that the peptide ratios are normally distributed; a demand for using most standard test statistics.

## Conclusion.

- **The SQL-based proteome data handling platform that we here present, solves the lack of multisample statistical analysis support in Matrixscience Mascot server and Distiller workpackage.**
- **Our solution presented here do require access to a workstation with Matrixscience Distiller v. 2.3 or 2.4**

## Key functions of Condenser

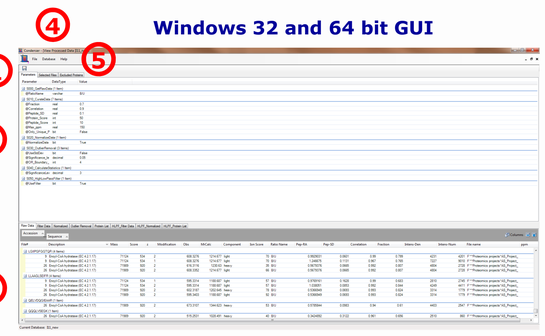
**1**



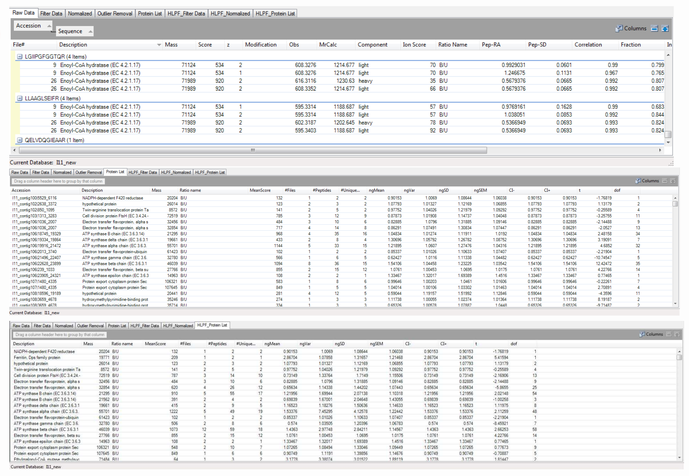

The processed data can be exported in a single step into a tabulated Excel format for easy visualization and data mining.

- Ratio vs. Intensity plots can visualize data distribution
- p-value vs. ratio plots can visualize significance at each ratio
- Density plots can visualize protein distribution

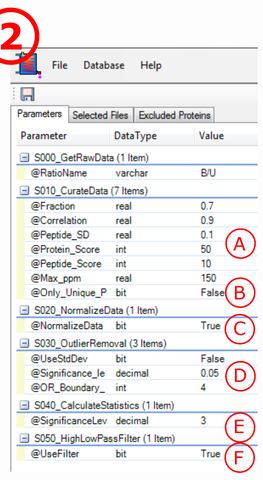
**4** **Windows 32 and 64 bit GUI** **5**



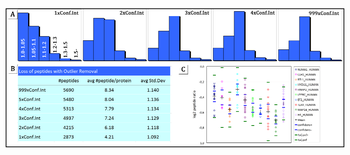

**3**



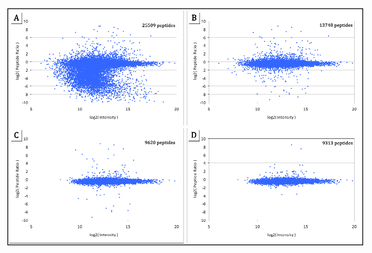
**2**




Raw data xml-files from Distiller are imported, combined and re-filtered based on user defined thresholds for the spectral quality values known from Distiller (A). The user has several options for choosing appropriate processing of the combined peptide list, including inclusion/exclusion of shared peptides (B), global normalization (C), outlier removal (D) and choice of significance threshold for t-test statistical calculations (E). In addition, the user can choose to apply a High-Low-Pass Filter (F) to correct for Distillers tendency to discard stable isotope labeled peptide doublets when differences in relative abundance exceed a certain magnitude.



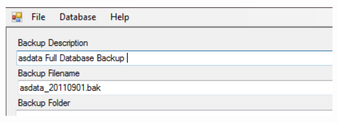

The effect of method used for outlier removal and the effect of the stringency chosen by the user. A) Histograms of the peptide standard errors with decreasing stringency of outlier removal are shown. B) The effect of stringent outlier removal can be seen as a decreasing number of peptides available for each protein quantization. C) Example of peptide outlier removal effect on a selection of ten proteins is shown.




Plots showing the effect of data filtering steps applied to peptide ratios before being condensed into a list of relative protein abundance. A) shows raw data before filtering is applied, B) shows data filtered using a mascot peptide score cutoff of 35. C) shows the same data where only peptide spectra with quantification features accounting for more than 70% of the peak area are included and where the correlation with the theoretical fitted peak envelope is more than 0.9. D) shows the same as C), but with the inclusion of a threshold of 0.3 for the standard error of ratio determination for each peptide.

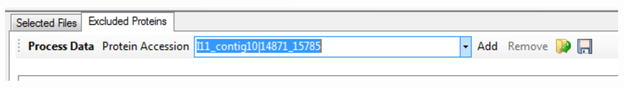
How did the data get from A to B? Condenser shows the data tables for each step in the user defined processing from the raw data in the imported Distiller xmls to the final spreadsheet-ready protein table.

**4**




Project databases can be backed up, restored and shared between collaborators using the database maintenance options.

**5**




The user can create a project specific exclusion list to define experimental protein contaminants for subsequent omission.

## References & Acknowledgements

**1.** Anders Dahl Knudsen, Henrik Kjeldahl, Daniel Erik Otzen and Allan Stensballe: Condenser: A statistical aggregation tool for integrating multi-sample quantitative proteomic data from Matrix Science's Mascot Distiller™; In preparation

**Acknowledgements** The authors are not affiliated in any way with Matrixscience but appreciate the versatility and transparency of the Distiller workpackage.